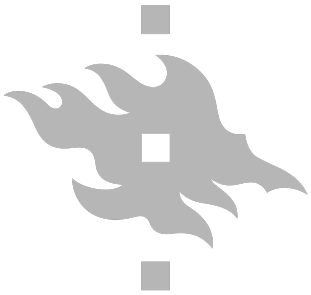


Helsingin Yliopisto
Valtiotieteellinen tiedekunta
Yhteiskuntatilastotiede
pro gradu -tutkielma

Moni-imputointi

Vastauskadon vaikutuksien korjaaminen kuluttajabarometriaineistossa

Mikko Patronen
Opiskelijanumero: 014492183

Toukokuu 2020

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta / Osasto — Fakultet / Sektion — Faculty

Valtiotieteellinen tiedekunta

Tekijä — Författare — Author

Mikko Patronen

Työn nimi — Arbetets titel — Title

Moni-imputointi: Vastauskadon vaikutuksien korjaaminen kuluttajabarometriaineistossa

Oppiaine — Läroämne — Subject

Yhteiskuntatilastotiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	Toukokuu 2020	31 s. + liitteet

Tiivistelmä — Referat — Abstract

Kato on yksi otanta-aineiston virhelähteistä. Se voi aiheuttaa aineistosta laskettaviin estimaatteihin harhaa, joten sen hallintaan on pyritty kehittämään erilaisia menetelmiä. Yksi tällainen menetelmä on imputointi, eli puuttuviksi jääneiden arvojen korvaaminen hyvin perustelluilla arvoilla. Estimointiin liittyvä epävarmuus tulee parhaiten huomioiduksi moni-imputoinnilla, mikä tarkoittaa useamman imputoidun aineiston muodostamista.

Tässä tutkielmassa perehdytään vastauskadon ominaisuuksiin. Imputointimenetelmän valintaan vaikuttaa esimerkiksi imputoitavan muuttujan asteikko sekä oletus kadon taustalla olevasta mekanismista. Imputoinnin apuna voidaan hyödyntää myös mahdollisesti käytössä olevia taustamuuttujia, jotka ovat yhteydessä imputoitavien muuttujien arvoihin ja niissä ilmenevään vastauskatoon. Myös tutkittavan ilmiön teorian kannalta olennaisia muuttujia voidaan hyödyntää.

Tutkielmassa tarkastellaan vuoden 2017 tammikuun Kuluttajabarometriaineistosta neljän kysymyksen osa-aineistoa, joka muodostaa kuluttajien luottamusindikaattorin. Kuluttajien luottamusindikaattori kuvaa 18-84 -vuotiaiden suomalaisten näkemyksiä ja odotuksia sekä henkilökohtaisesta että Suomen yleisestä taloustilanteesta. Kiinnostuksen kohteena on erityisesti selvittää, vääristääkö vastauskato aineistosta laskettavia estimaatteja.

Tutkielmassa vastauskatoa paikataan moni-imputoimalla käyttäen hot deck -imputointia, jossa puuttuvat tiedot korvataan taustatiedoiltaan mahdollisimman samankaltaisilta vastaajilta kopioiduilla arvoilla. Työssä muodostetaan viisi imputointimallia käyttäen erilaisia yhdistelmiä taustamuuttujista. Taustatieto ikäluokasta osoittautuu tärkeäksi mallimuuttujaksi tulosten kannalta. Imputointimalli ilman ikäluokkatietoa pienentää luottamusindikaattorin estimaattia sekä koko aineiston tasolla että sukupuoliryhmissä. Luottamusindikaattorin arvot estimoituvat alkuperäisen aineiston estimaattia pienemmiksi myös, jos malli perustuu ainoastaan tietoon sukupuolesta.

Avainsanat — Nyckelord — Keywords

vastauskato, tilastomenetelmät, imputointi, estimointi

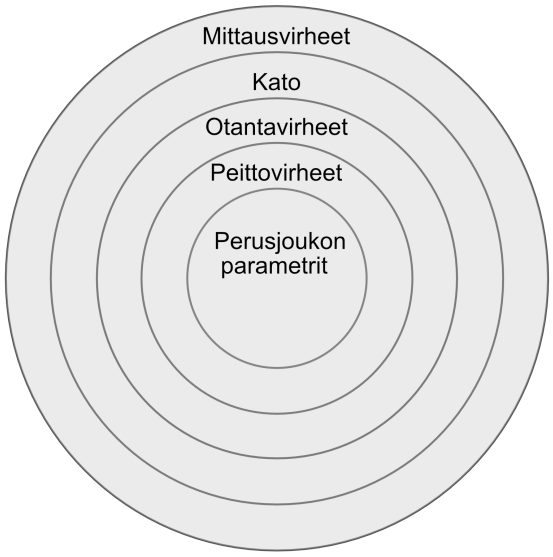
Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Sisältö

1 Johdanto

Tilastollisten analyysien ja tunnuslukujen laskemisen tarkoituksena on esimerkiksi saada konkreettista näyttöä asioiden tilasta. Monenlaisissa tutkimustilanteissa olisi kätevää, jos tilastollisen tiedonkeräyksen ja mittauksen voisi suorittaa kaikille tutkittavan ilmiön kannalta olennaisille tutkimusyksiköille, mutta tämä on käytännössä usein taloudellisesti tai fyysisesti mahdotonta toteuttaa. Esimerkiksi autonrengasvalmistajan olisi mahdotonta testata jokaista myyntiin toimitettavaa rengasta ilmoittaakseen ennusteen sille, kuinka monta ajokilometriä renkaat tulevat kestämään käytössä. Sen sijaan yleensä tutkimuksen perusjoukkoa ja sen ominaisuuksia koskeva päättely suoritetaan keräämällä perusjoukosta otos, joka oikein poimittuna edustaa perusjoukon ominaisuuksia oikeassa suhteessa perusjoukkoon nähden. Tällä tavoin perusjoukon parametreja koskeva tutkimus tehdään niin sanotulle otanta-aineistolle. Otanta-aineistolle suoritettava tutkimus altistuu tutkimuksen edetessä erilaisille virhelähteille, jotka vaikuttavat aineiston perusteella tehtävien päätelmien luotettavuuteen ja tarkkuuteen. Mainittuja virhelähteitä ovat peittovirheet, otantavirheet, kato ja mittausvirheet. Alwinin (2007) visualisoima kehärakenne kuvaa virhelähteiden rakennetta hyvin (kuva 1.1).



Kuva 1.1: Otanta-aineiston virhelähteet. Kuva muokattu alkuperäisestä lähteestä (Alwin, 2007)

Yhteiskunnassa tilastollista tietoa voidaan käyttää päätöksenteon tukena. Jos halutaan päätösten perustuvan mahdollisimman tarkkaan tietoon, analyysien ja tunnuslukujen laskemiseen tulee käyttää menetelmiä, jotka tuottavat mahdollisimman tarkkaan todellisuutta vastaavia tuloksia.

Tilastokeskus kerää kuukausittain haastattelututkimuksena kuluttajabarometria, joka mittaa vastaajien ajatuksia ja ennustuksia sekä henkilökohtaisen talouden, että Suomen talouden tilasta. Kuluttajabarometrin aineistosta laskettava kuluttajien luottamusindikaattori ennustaa hyvin kuluttajien talouskäyttäytymistä, joten sitä käytetään poliittiseen päätöksentekoon.

Mielipiteitä ja uskomuksia mittaavia kyselyitä voidaan sensitiivisyytensä vuoksi pitää otollisina tuottamaan katoa (Durrant, 2005). Kuluttajabarometrin vastausprosentti on noin 55 prosenttia, eli tässä tapauksessa 45 prosenttia vastaajista jättää kokonaan vastaamatta. Kadon vaikutuksia aineistosta saataviin tuloksiin voidaan pyrkiä hallitsemaan erilaisin painoin tai paikkaamalla puuttuvat rivit jollakin hyvin perustellulla menetelmällä. Ne henkilöt, jotka ovat vastanneet joihinkin, mutta eivät kaikkiin kysymyksiin, jättävät aineistoon yksittäisiä puuttuvia tietoja. Myös tällaista katoa voidaan pyrkiä hallitsemaan korvaamalla puuttuneiksi jääneet arvot jollakin hyvin perustellulla menetelmällä. Puuttuvien vastausten korvaaminen muodostaa paikatun, eli imputoidun aineiston, jolle voidaan suorittaa halutut tarkastelut ja analyysit. Muodostamalla useampia paikattuja aineistoja puhutaan moni-imputoinnista.

Tässä tutkielmassa tarkastellaan kadon vaikutuksien hallintaa moni-imputoinnilla kuluttajabarometrin kuluttajien luottamusindikaattoria mittaavissa muuttujissa. Tarkastelun kohteena ovat demografisten luokkien erot kuluttajien luottamusindikaattorin estimaattien keskivirheissä lähtötilanteessa ja imputoidussa aineistossa.

Tämän tutkielman aiheen kannalta olennainen otanta-aineiston virhelähde on

kato. Luvussa 2 esitellään kadon ja sen hallitsemisen teoriaa. Luvussa 3 esitellään kuluttajabarometrin aineisto, luottamusindikaattorin estimaatit ja vastauskadon esiintyminen demografisten tekijöiden luokissa. Luvussa 4 muodostetaan viisi erilaista imputointimallia ja muodostetaan niihin liittyviä hypoteeseja. Moni-imputoiduista aineistoista lasketaan estimaatit ja vertaillaan eri mallien tuloksia. Luvussa 5 tiivistetään ja pohditaan tuloksia. Luku 6 kertaa työn vaiheet.

2 Kato

Kato on yksi otanta-aineiston virhelähteistä. Sillä viitataan tietoon, joka oli tarkoitettu kerättäväksi, mutta joka jäi jostakin syystä puuttuvaksi. Kato voidaan jakaa yksikkövastauskatoon ja erävastauskatoon (kuva 2.1). Yksikkövastauskadolla tarkoitetaan tilannetta, jossa yksiköltä ei saada kerättyä mitään tietoa (Little & Rubin, 2002). Tätä tilannetta voidaan pyrkiä korjaamaan esimerkiksi erilaisia painoja käyttämällä tai korvaamalla puuttuvia arvoja moni-imputoinnilla. Erävastauskadolla tarkoitetaan tilannetta, jossa tilastoyksiköltä (henkilö, yritys, sensoritieto yms.) kerättävä tieto jää jostakin syystä pois aineistosta (Little & Rubin, 2002). Yksinkertainen esimerkki erävastauskadosta ilmenee, kun henkilö on haluton tai osaamaton vastaamaan esitettyyn kysymykseen ja kysytty tieto jää keräämättä.

Kuva 2.1: Kadon jaottelu

Pelkän vastaamattomuuden lisäksi erävastauskatoa saattaa syntyä esimerkiksi teknisen virheen tai inhimillisen erheen vuoksi. Tekninen virhe voi tarkoittaa esimerkiksi tilannetta, jossa sensori jättää rekisteröimättä havainnon, ja inhimillinen virhe tilannetta, joka syntyy esimerkiksi tahattomasta tai tahallisesta

virheellisestä vastauksesta tai vastauksen kirjaamisesta.

Erävastauskato vähentää käytettävän tiedon määrää ja haittaa aineiston välitöntä käyttöä datan analysointiin. Tunnuslukujen laskeminen vain saatavilla olevasta datasta jättää huomioimatta mahdollisuuden, että vastanneet ja vastaamattomat henkilöt eroavat jonkin vastauksiin vaikuttavan tekijän suhteen systemaattisesti. Tämä aiheuttaa tuloksiin harhaa. Erävastauskadon vaikutuksia voidaan pyrkiä hallitsemaan korvaamalla puuttuneiksi jääneet arvot erilaisilla imputointimenetelmillä.

2.1 Kadon mekanismit

Tässä tutkielmassa tarkastellaan vastauskadon hallintaa moni-imputoinnin avulla aineistolla, joka sisältää sekä yksikkö- että erävastauskatoa. Aineiston perusteella suoritettava validi päättely edellyttää kadon taustalla olevan mekanismin huomioimista. Mekanismi tulkitaan sen perusteella, mitä tiedetään tutkittavan muuttujan ja muiden muuttujien välisestä suhteesta.

Määritellään kokonainen kadoton $(n \times K)$ aineisto $Y = (y_{ij})$, missä i. rivi $y_i = (y_{i1}, ..., y_{iK})$ sisältää arvot y_{ij} muuttujissa Y_j yksiköille i. Olkoon katoa sisältävä kadon indikaattorimatriisi $M = M_{ij}$, missä $m_{ij} = 1$, jos solun y_{ij} arvo puuttuu, ja $m_{ij} = 0$, mikäli solulla y_{ij} on havaittu arvo. Kadon mekanismin määrittää ehdollinen jakauma M ehdolla Y, esimerkiksi $f(M|Y, \phi)$, missä ϕ edustaa tuntemattomia parametreja. Jos

$$f(M|Y, \phi) = f(M|\phi) \quad \text{kaikille } Y, \phi, \tag{2.1}$$

puuttuvan tiedon ilmenemistodennäköisyys on täysin riippumatonta kyseisten muuttujien tai muiden muuttujien arvoista, on kato täysin satunnaista, MCAR (missing completely at random) (Little & Rubin, 2002). Surveytutkimuksessa tämä tilanne toteutuisi, jos puuttuvat tiedot olisivat täysin riippumattomia

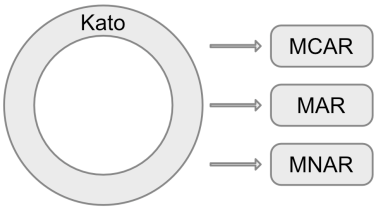
puuttuvia tietoja sisältävistä muuttujista, vastaajien taustatiedoista ja muista muuttujista. Täysin satunnainen kato on käytännössä harvinaista ja etenkin sosiaalitieteiden alan tutkimuksissa MCAR-oletus on liian voimakas ja jää todennäköisesti täyttymättä (Durrant, 2005).

Merkitään aineiston Y havaittuja arvoja Y_{obs} ja puuttuneita arvoja Y_{mis}. Jos puuttuvien tietojen ilmenemistodennäköisyydet riippuvat havaitusta aineistosta, mutta eivät puuttuvista arvoista, siis

$$f(M|Y, \phi) = f(M|Y_{obs}, \phi) \quad \text{kaikille } Y_{mis}, \phi, \tag{2.2}$$

sanotaan, että kato satunnaista, MAR (missing at random). Surveytutkimuksessa tämä tilanne tarkoittaisi, että esimerkiksi vastaajan tuloluokka tai ikä selittää vastaamattomuutta. Surveyaineistoissa vastanneiden ja vastaamattomien välillä voidaan olettaa olevan jokin systemaattinen erottava tekijä (Rubin, 2004).

Kato ei ole satunnaista, NMAR (not missing at random), jos M:n jakauma riippuu aineiston Y puuttuvista arvoista. Kun puuttuvien tietojen ilmenemistodennäköisyys riippuu muuttujasta itsestään, eikä ehdollistaminen muihin muuttujiin muuta tilannetta. NMAR on käytännössä hankala osoittaa. Tässä esitelty kadon mekanismien määrittely ja jaottelu noudattaa Rubinin ja Littlen (2002) määritelmiä (kuva 2.2).



Kuva 2.2: Kadon mekanismit

Edellä mainitut kadon mekanismit voidaan edelleen jakaa kahteen ryhmään (kuva 2.3). MCAR ja MAR voidaan tulkita sivuutettavissa oleviksi mekanis-

meiksi, mikä tarkoittaa, että niitä voidaan hallita. MNAR puolestaan on ei-sivuutettavissa oleva. Yleensä analyysimenetelmien edellytyksenä on oletus sivuutettavissa olevasta katomekanismista.

Kuva 2.3: Kadon mekanismien jaottelu

2.2 Imputointimenetelmiä

Sekä yksikkö- että erävastauskadon vaikutuksia voidaan pyrkiä hallitsemaan erilaisilla imputointimenetelmillä, eli korvaamalla puuttuvat havainnot hyvin perustelluilla arvoilla. Imputointimenetelmä voi perustua malliluovuttajaan (model-donor imputation) tai vastaajaluovuttajaan (real-donor imputation). Malliluovuttaja-imputointimenetelmässä puuttuvat arvot korvataan tilastollisen mallin avulla ja vastaajaluovuttaja-imputointimenetelmässä puuttuvat arvot korvataan ominaisuuksiltaan samankaltaiselta vastaajalta (Laaksonen, 2013). Lopputuloksena syntyy yksi tai useampi paikattu aineisto, jolle voidaan suorittaa halutut tarkastelut ja analyysit.

Yksinkertaisimmillaan puuttuvat luvut korvataan muuttujan keskiarvolla tai mediaanilla. Tämä menetelmä lisää käytettävissä olevien havaintojen määrää, mutta aliarvioi muuttujan todellista varianssia. Ennen kaikkea tämä menetelmä olettaa, että vastaamattomien keskiarvo tai mediaani on sama vastanneiden kanssa, joten sen käyttö ei ole suositeltavaa (Rubin, 2004). Hot deck -imputoinniksi kutsutaan menetelmää, jossa puuttuva arvo korvataan samasta aineistosta taustatiedoiltaan samankaltaiselta vastaajalta poimitulla arvolla (Rubin, 2004). Cold deck -imputoinnissa korvaava arvo saadaan mahdollisesti aiemmin kerätystä tiedosta samalta vastaajalta (Rubin, 2004). Hot deck- ja

cold deck -imputoinnit ovat vastaajaluovuttaja-imputointimenetelmiä.

Yksinkertaisessa regressioimputoinnissa luodaan malli, jonka sovite tuottaa selittävien muuttujien avulla puuttuvien arvojen tilalle korvaavat arvot (Little & Rubin, 2002). Menetelmässä korostuu apumuuttujan ja imputoitavan muuttujan välinen korrelaatio ja todellinen vaihtelu jää aliarvioiduksi. Stokastisessa regressioimputoinnissa malli saadaan paremmin kuvaamaan todellista vaihtelua lisäämällä edellä mainittuun regressiosovitteeseen satunnainen virhetermi $\hat{e}$ (Little & Rubin, 2002).

2.3 Moni-imputointi

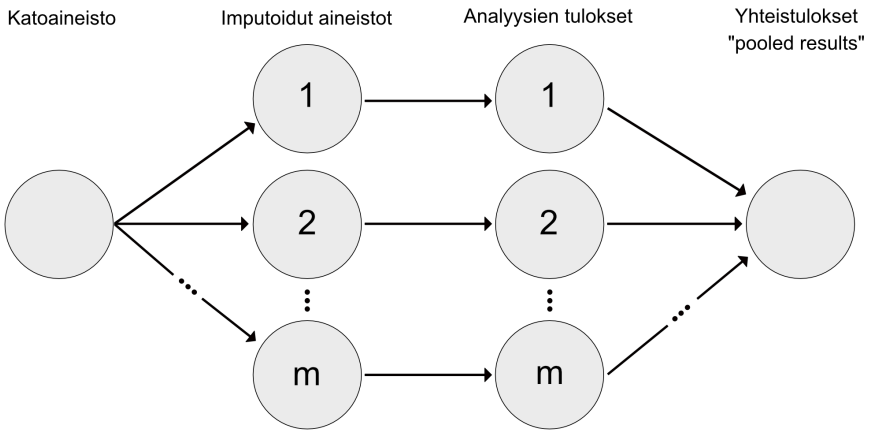
Puuttuvien arvojen aiheuttamaa epävarmuutta mallinnetaan yksittäistä imputointikertaa paremmin suorittamalla imputointi samalle aineistolle useaan kertaan. Tätä menetelmää nimitetään moni-imputoinniksi.

Katoa sisältävän aineiston paikkaus moni-imputoinnilla koostuu kolmesta vaiheesta (van Buuren, 2018):

1. Imputointivaihe
2. Analyysivaihe
3. Yhdistämisvaihe

Imputointivaiheessa imputoituja aineistoja luodaan useampia; tilanteesta riippuen 2 - 20 kappaletta. Määrän valintaan vaikuttavat aineiston koko sekä kadon prosenttiosuus aineistossa (Rubin, 2004). Imputointiin voidaan käyttää esimerkiksi stokastista regressioimputointia M kertaa luomaan M kappaletta imputoituja aineistoja. Jokainen luotu aineisto simuloi omalta osaltaan muuttujien hajontaan liittyvää epävarmuutta, jolloin useamman aineiston yhdistäminen mallintaa hyvin todellista hajontaan liittyvää epävarmuutta.

Analyysivaiheessa jokaiselle imputoidulle aineistolle *1, 2,..., M* suoritetaan halutut tilastolliset analyysit. Yhdistämisvaiheessa erillisten aineistoanalyysien tulokset tai parametrien estimaatit yhdistetään. Moni-imputoinnin vaiheet on esitelty kaaviona kuvassa 2.4.



Kuva 2.4: Moni-imputoinnin vaiheet. Kuva muokattu alkuperäisestä lähteestä (van Buuren, 2018)

Moni-imputoidun aineiston parametrin piste-estimaatti on imputoitujen aineistojen estimaattien keskiarvo. Moni-imputoidun aineiston varianssin estimaatti (yhtälö 2.5) lasketaan ottaen huomioon imputoitujen aineistojen sisäinen (yhtälö 2.3) ja aineistojen välinen varianssi (yhtälö 2.4).

$$Var_{within} = \frac{\sum_{i=1}^{M} SE_i^2}{M} \tag{2.3}$$

$$Var_{between} = \frac{\sum_{i=1}^{M} (\beta_i - \bar{\beta})^2}{M - 1} \tag{2.4}$$

$$Var_{total} = Var_{within} + Var_{between} + \frac{Var_{between}}{M} \tag{2.5}$$

Kaavoissa Var_{within} = aineistojen sisäinen varianssi, $Var_{between}$ = aineistojen välinen varianssi, SE = keskivirhe, β = parametriestimaatti, M = imputoitujen aineistojen lukumäärä ja Var_{total} = kokonaisvarianssi (Rubin, 2004).

Yleisesti moni-imputointiin käytettyjä menetelmiä ovat EM (expectation-maximization) -algoritmi sekä Markov Chain Monte Carlo -menetelmä, jossa oletetusta yhteisjakaumasta simuloidaan puuttuvien havaintojen korvaavat arvot.

3 Aineisto

Tutkielman tarkastelun kohteena on Tilastokeskuksen keräämä kuluttajabarometri ja sen muuttujista laskettava kuluttajien luottamusindikaattori. Kuluttajabarometrin tuloksia käytetään Suomessa poliittiseen päätöksentekoon, sillä ne ennustavat hyvin kuluttajien talouskäyttäytymistä. Toukokuusta 2019 alkaen tutkimuksen nimi on ollut kuluttajien luottamustutkimus, mutta tässä tutkielmassa käytetään aineiston nimitystä sen keräysvuodelta 2017 (Tilastokeskus, 2017). Vuodesta 2019 alkaen aineistoa on kerätty puhelinhaastattelulla ohella myös internetlomakkeella, jolloin aineiston analyysissä tulee tiedonkeruutapojen vuoksi ottaa huomioon myös moodiefekti. Tämä hankaloittaisi tutkielman aiheen selkeää käsittelyä, joten aineisto on valittu ajanjaksolta ennen internetkyselyiden aloittamista.

Barometrin tutkimusalueena on koko Suomi ja sen perusjoukkoon kuuluu 4,5 miljoonaa henkilöä 2,6 miljoonasta kotitaloudesta. Otokseen päätyneet edustavat Suomen 15 - 84 -vuotiaita iän, sukupuolen, asuinalueen ja äidinkielen suhteen. Brutto-otoskoko on 2 350 henkilöä ja se on poimittu Tilastokeskuksen väestötietokannasta systemaattisella satunnaisotannalla. Otoskehikko on lajiteltu siten, että se noudattaa maantieteellistä väestötiheyttä. Aineiston yksikkövastauskato on nykyään yli 45 prosenttia, jolloin vastauksia saadaan kuukausittain noin 1 200 henkilöltä. Katoon sisältyy tutkimuksesta kieltäytyneiden lisäksi ne otokseen valitut henkilöt, joita ei tavoitettu.

Tutkielmassa tarkasteltavan kuluttajien luottamusindikaattorin neljä osateki-

jää ovat kuluttajabarometrin kyselyn kysymykset:

- K1: Millaisen arvioitte Suomen taloudellisen tilanteen olevan 12 kuukauden kuluttua verrattuna tilanteeseen nyt?
- K2: Millainen on oman kotitaloutenne taloudellinen tilanne nyt verrattuna tilanteeseen 12 kuukautta sitten?
- K3: Millaisen arvioitte kotitaloutenne taloudellisen tilanteen olevan 12 kuukauden kuluttua verrattuna tilanteeseen nyt?
- K4: Verrattuna edelliseen 12 kuukauteen, miten aiot käyttää rahaa kestokulutustavaroiden hankintaan seuraavan 12 kuukauden aikana?

Haastateltava arvioi kysymyksien K1, K2 ja K3 vastauksia vaihtoehdoista "Paljon parempi", "Jonkin verran parempi", "Samanlainen", "Jonkin verran huonompi", "Paljon huonompi" ja kysymyksen K4 vastausta vaihtoehdoista "Paljon enemmän", "Jonkin verran enemmän", "Saman verran", "Jonkin verran vähemmän", "Paljon vähemmän". Vastausvaihtoehdot on numeroitu edellisen mukaisessa järjestyksessä 1, 2, 3, 4, 5.

Kaikki tässä tutkielmassa mainitut aineiston käsittelyt on toteutettu SAS Enterprise Guide 7.1 -ohjelmistolla. Aineiston visualisoinnit on toteutettu R Studio -ohjelmistolla ggplot2-pakettia käyttäen. Luottamusindikaattorin estimaatit ja keskivirheet laskettiin PROC SURVEYMEANS -komennolla.

3.1 Luottamusindikaattorin estimaatit ja keskivirheet

Käsittelyvaiheessa muuttujien K1, K2, K3 ja K4 arvot skaalataan niin, että skaalatut arvot vaihtelevat -100:n ja +100:n välillä. Skaalauksessa asteikon arvo 5 skaalautuu arvoon -100 ja arvo 1 arvoon 100. Luottamusindikaattorin arvo on skaalattujen saldolukujen keskiarvo. Korkeampi luottamusindikaattorin lukema kertoo valoisammasta näkemyksestä taloudesta.

Taulukosta 3.1 voidaan havaita, että koko aineiston luottamusindikaattorin arvo on noin 5,1. Miesten näkemys taloudesta on positiivisempi (noin 7,5) naisiin verrattuna (noin 2,8). Sekä miesten että naisten ryhmissä luottamusindikaattorin arvo laskee ikäluokan kasvaessa. Koulutusasteen kasvaessa molemmissa sukupuoliryhmissä luottamusindikaattorin estimaatti kehittyy siten, että indikaattorin arvo kasvaa koulutusasteen kasvaessa sillä poikkeuksella, että indikaattori saa pienimmän arvonsa molemmissa sukupuoliryhmissä keskiasteen koulutusluokassa.

Taulukko 3.1
Luottamusindikaattorin estimaatit ja keskivirheet sukupuolen ja ikäluokan mukaan

		Estimaatti	Keskivirhe	Variaatiokerroin
Koko aineisto		5,137	0,716	0,139
Miehet		7,494	1,019	0,136
Ikäluokka	alle 35	11,015	2,124	0,192
	35-49	8,57	2,082	0,242
	50-64	5,114	1,826	0,357
	yli 65	3,563	1,768	0,490
Koulutus	Perusaste	7,757	1,948	0,251
	Keskiaste	5,031	1,618	0,322
	Ammattiopisto, AMK	9,743	2,111	0,217
	Korkeakoulututkinto, ylempi	12,558	2,864	0,228
Naiset		2,823	0,999	0,353
Ikäluokka	alle 35	10,391	1,960	0,189
	35-49	4,597	2,423	0,527
	50-64	-2,373	1,891	-0,797
	yli 65	-1,857	1,538	-0,828
Koulutus	Perusaste	2,600	2,003	0,770
	Keskiaste	0,781	1,770	2,265
	Ammattiopisto, AMK	4,241	2,061	0,486
	Korkeakoulututkinto, ylempi	6,199	1,874	0,302

3.2 Vastauskadon esiintyminen aineistossa

Taulukosta 3.2 selviää, että puuttuvien havaintojen määrä luottamusindikaattorin muuttujissa on lähes identtinen ja että vastauskatoprosentti on suuri, noin 48 prosenttia.

Taulukko 3.2
Vastauskato luottamusindikaattorin muuttujissa

Muuttuja	Puuttuvia vastauksia	Katoprosentti
K1	1132	48,17
K2	1132	48,17
K3	1132	48,17
K4	1133	48,21

Miesten vastauskatoprosentti on noin 49 prosenttia ja naisten vastauskatoprosentti noin 47 prosenttia. Sukupuolten välillä ei siis ilmene suurta eroa vastauskatoprosenteissa (taulukko 3.3). Molemmissa sukupuoliryhmissä vastauskatoprosentti pienenee ikäryhmittäin iän kasvaessa sekä koulutusasteen kasvaessa.

Taulukko 3.3
Vastauskatoprosentit taustamuuttujissa

Muuttuja		Otos	Vastanneet	Katoprosentti
Koko aineisto		2350	1217	48,21
Miehet		1135	574	49,43
Ikäluokka	alle 35	356	145	59,27
	35-49	265	130	50,94
	50-64	291	163	43,99
	yli 65	223	137	38,57
Koulutus	Perusaste	341	136	60,12
	Keskiaste	517	249	51,84
	Ammattiopisto, AMK	198	131	33,84
	Korkeakoulututkinto, ylempi	79	59	25,32
Naiset		1215	643	47,08
Ikäluokka	alle 35	340	142	58,24
	35-49	265	126	52,45
	50-64	303	182	39,93
	yli 65	307	193	37,13
Koulutus	Perusaste	345	149	56,81
	Keskiaste	458	246	46,29
	Ammattiopisto, AMK	273	155	43,22
	Korkeakoulututkinto, ylempi	139	93	33,09

4 Moni-imputoinnin soveltaminen aineistoon

Tässä luvussa moni-imputoidaan luottamusindikaattorin muuttujien puuttuvat vastaukset erilaisilla imputointimalleilla. Mallin valinnassa on huomioitava, että tässä aineistossa imputoitavat muuttujat ovat järjestysasteikollisia muuttujia ja niiden muunnos luottamusindikaattoriksi edellyttää kokonaislukujen käyttöä. Mallipohjaisten imputointitulosten pyöristämisen on osoitettu tuottavan mahdollisesti harhaisia tuloksia (Allison, 2005).

Hyvä tapa täyttää kokonaislukuvaatimus on korvata puuttuvat tiedot hot deck -imputoinnilla, eli vastaajaluovuttaja-menetelmällä. Vastaajaluovuttaja-imputoinnissa puuttuva arvo korvataan aidosti havaitulla arvolla taustatiedoiltaan mahdollisimman samankaltaiselta vastaajalta (Laaksonen, 2013). Tilastokeskuksen väestötietokannasta poimittuna kyselyyn vastaamattomien taustatiedot ovat tiedossa otosaineistosta, joten vastaajaluovuttaja-imputoinnin soveltamiselle on hyvät edellytykset. Tässä aineistossa mukana olevat taustatiedot ovat sukupuoli-, ikä-, koulutus-, siviilisääty- ja maakuntatiedot. Tuloksia tarkastellaan koko aineiston tasolla ja sukupuolten välillä ikäryhmissä.

4.1 Taustamuuttujien yhteys vastauskatoon ja luottamusindikaattorin arvoon

Van Buurenin, Boshuizenin ja Knookin (1999) mukaan imputointimalliin suositellaan valittavaksi kaikki aineistosta saatavilla olevat muuttujat, jotka

1. Kuuluvat tutkimuksen teoriaan
2. Ovat yhteydessä vastauskatoon imputoitavissa muuttujissa
3. Ovat yhteydessä imputoitavien muuttujien arvoihin

Taustamuuttujien tilastollista yhteyttä vastauskatoon tarkasteltiin χ^2 -riippumattomuustestillä. Taustamuuttujien yhteys luottamusindikaatorin

arvoon selvitettiin kaksiluokkaisissa muuttujissa riippumattomien parien t-testillä tai muissa muuttujissa varianssianalyysillä. χ^2 -riippumattomuustestiä varten luotiin dummy-muuttuja indikoimaan puuttuvia vastauksia.

Tulosten perusteella tieto vastaajan maakunnasta ei ollut yhteydessä vastauskatoon eikä luottamusindikaattorin arvoon, joten sitä ei otettu mukaan imputointimalliin. Tämän lisäksi mallista jätettiin pois siviilisääty-muuttuja. Sen yhteys osoittautui tilastollisesti merkitseväksi sekä vastauskatoon että luottamusindikaattorin arvoon, mutta sen käyttö olisi pienentänyt vastaajaluovuttajaluokkien kokoja useissa ryhmissä hyvin pieniksi. Yleinen ratkaisu tällaisessa tilanteessa on yhdistää luokkia, mutta tässä tapauksessa yhdistämisessä ongelmaksi olisi muodostunut päättää esimerkiksi kysymys siitä, liitetäänkö lesket sinkkuihin vai naimisissa oleviin. Tämän ongelman kiertämiseksi muuttuja jätettiin pois imputointimallista.

Ikäluokka ja koulutus ovat tilastollisesti merkitsevästi yhteydessä sekä vastauskatoon että luottamusindikaattorin arvoon (taulukko 4.1). Lisäksi sukupuoli on tilastollisesti yhteydessä luottamusindikaattorin arvoon.

Taulukko 4.1
Taustamuuttujien yhteys vastauskatoon ja luottamusindikaattorin arvoon

	Vastauskato		Luottamusindikaattorin arvo	
Muuttuja	$\chi^2(p)$	Cramer's V	t-testi (p)	ANOVA (p)
Sukupuoli	1,20 (0,27)	-0,0226	3.65 (<0.001)	
Ikäluokka	66,34 (<0,001)	0,168		12,66 (<0.001)
Koulutus	72,38 (<0,001)	0,176		2,70 (0,044)

4.2 Imputointimallin muodostaminen

Edellisen tarkastelun mukaan ikäluokka ja koulutusaste ovat tilastollisesti merkitsevästi yhteydessä sekä vastauskatoon että luottamusindikaattorin arvoon, joten ne valitaan mukaan imputointimalliin. Tämän lisäksi sukupuoli on tilastollisesti merkitsevästi yhteydessä luottamusindikaattorin arvoon. Vaikka tämä

muuttuja ei ole tilastollisesti merkitsevästi yhteydessä vastauskatoon, niin se lisätään mukaan imputointimallin käytössä oleviin muuttujiin. Taulukon 3.1 tuloksista nähdään isoja eroja sukupuolten välillä, joten sitä voidaan pitää olennaisena muuttujana mallin kannalta. Koska kyseessä on surveyaineisto, voidaan imputointimallia muodostettaessa olettaa taustamuuttujien vaikuttavan vastauskatoon. Tämä tarkoittaa, että kadon mekanismin voidaan olettaa olevan satunnainen (MAR) (Little Rubin, 2002).

Koska vastauskato on suuri (yli 48 prosenttia) valitaan imputointien lukumääräksi 20. Moni-imputointi toteutetaan PROC SURVEYIMPUTE-komennolla. Jokaiselle imputoidulle aineistolle lasketaan estimaatit ja estimaattien keskivirheet PROC SURVEYMEANS -komennolla. Näiden 20 datatiedoston estimaatit ja niiden keskivirheet edustavat yhdessä vastauskadon muodostamaa epävarmuutta muuttujissa. Koonti suoritetaan PROC MIANALYZE -komennolla, joka yhdistää imputoitujen data-aineistojen tulokset yhdeksi koko aineistoa edustavaksi luottamusindikaattorin estimaatiksi ja sen keskivirheeksi.

4.3 Imputointimallien hypoteesit

Seuraavaksi muodostetaan erilaisia imputointimalleja taustamuuttujien erilaisilla kombinaatioilla. Hot deck -imputointimallissa määritellään taustamuuttujaryhmät, joiden sisällä malliluovuttaminen tapahtuu. Luotuja malleja voidaan suhteuttaa sivun 13 periaatteisiin imputointimallin muodostamisesta. Asetan ennen estimaattien laskuja hypoteesit jokaiselle mallille tulosten suunnasta suhteessa alkuperäiseen aineistoon. Muodostettavat imputointimallit ovat:

- Malli 1: Sukupuoli
- Malli 2: Sukupuoli ja ikäluokka
- Malli 3: Sukupuoli ja koulutus
- Malli 4: Ikäluokka ja koulutus
- Malli 5: Sukupuoli, ikäluokka ja koulutus

Malli 1: Tässä mallissa vastaajaluovuttaja-imputointi kerää korvaavat arvot perustuen vain tietoon sukupuolesta. Mieheltä puuttuvat arvot korvataan satunnaiselta mieheltä arvotulta luvulla. Taulukon 3.1 perusteella luottamusindikaattorin keskiarvo on miehillä 7,5 ja naisilla 2,8. Nuorimmassa ikäluokassa estimaatit ovat miehillä 11 ja naisilla 10,4, joista estimaatit laskevat vanhimpiin ikäluokkiin 3,6:een miehillä ja -1,9 naisilla. Taulukosta 3.3 käy ilmi, että vastauskato-osuus on korkein nuorimmassa ikäryhmässä, josta se pienenee vanhimpaan ikäryhmään saakka. Eniten vastauksia puuttuu molemmissa sukupolvissa ryhmässä, jonka luottamusindikaattorin estimaatti on korkein. Vastaavasti vastaajaluovuttajia on eniten ryhmistä, joissa luottamusindikaattori on matalampi. Tämä pätee molempien sukupuolten tilanteeseen.

Koko aineiston tasolla nuorempien ikäryhmien suurempi kato-osuus korvaantuu vanhempien ikäryhmien matalammilla luottamusindikaattorin arvoilla, jolloin koko aineiston keskiarvo tulee pienenemään. Ikäluokissa keskivirheet saattavat kasvaa, koska esimerkiksi nuorten ryhmään voi olettaa tulevan vaihtelua kasvattavia arvoja vanhemmista ikäluokista. Koko aineiston tasolla keskivirheen voi olettaa pienenevän, koska imputoitavat arvot kasvattavat havaintojen määrää ja toisaalta arvot kopioituvat jo valmiiksi aineistossa olevista havainnoista. Tämä peruste pätee muihinkin malleihin.

Mallin 1 hypoteesi: malli siirtää jokaisen ikäryhmän estimaattia kohti sukupuolen keskiarvoa ja koko aineiston estimaatti tulee pienenemään. Keskivirheet pienenevät.

Malli 2: Tässä mallissa vastaajaluovuttaja-imputointi perustuu tietoon sukupuolesta ja ikäluokasta. Vastauksia puuttuu ja näin ollen myös korvaantuu eniten nuoremmissa ikäluokissa. Koska korvaavat arvot tulevat samasta ryhmästä, niin aineistoon tulee suhteellisesti enemmän lisää korkeampia arvoja. Tämä nostanee hiukan koko aineiston keskiarvoa. Ikäluokittain oletan tulosten pysyvän samankaltaisina kuin ikäluokissa taulukossa 3.1. Keskivirheet piene-

nevän hieman.

Mallin 2 hypoteesi: mallin ikäryhmien estimaatit pysyvät alkuperäisen aineiston kaltaisina sukupuolittain, mutta koko aineiston estimaatti tulee kasvamaan hieman. Keskivirheet pienenevät hieman.

Malli 3: Tässä mallissa vastaajaluovuttaja-imputointi perustuu taustatietoihin sukupuolesta ja koulutuksesta. Taulukon 3.3 mukaan vastauskato pienenee koulutuksen lisääntyessä. Taulukon 3.1 mukaan luottamusindikaattorin arvo on suurempi korkeammin koulutetuilla. Vastauksia puuttuu ja siten myös korvaantuu eniten alemmin koulutetuissa ryhmissä. Koska korvaavat arvot tulevat samasta ryhmästä, niin aineistoon tulee suhteellisesti enemmän lisää pienempiä arvoja. Tämän voi olettaa laskevan hiukan koko aineiston keskiarvoa. Koulutusluokittain oletan tulosten pysyvän samankaltaisina kuin koulutusluokkien taulukossa 3.1. Keskivirheiden oletan pienenevän hieman.

Mallin 3 hypoteesi: mallin koulutusluokkien estimaatit pysyvät alkuperäisen aineiston kaltaisina sukupuolittain, mutta koko aineiston estimaatti tulee pienenemään hieman. Keskivirheet pienenevät hieman.

Malli 4: Tässä mallissa vastaajaluovuttaja-imputointi perustuu taustatietoihin ikäluokasta ja koulutuksesta. Tämä malli on hankala ennustettava, sillä luottamusindikaattorin arvo laskee ikäryhmittäin samalla kuin vastauskatoprosenttikin. Toisaalta ikä on ainakin osittain yhteydessä koulutusasteeseen, jonka noustessa myös luottamusindikaattorin arvo kasvaa. Nuorimmissa ikäluokissa voi olettaa olevan suhteellisesti eniten myös alhaisen koulutuksen ryhmiin kuuluvia. Tässä yhdistyy kahdella tavalla korkea vastauskato-osuus. Luottamusindikaattorin arvoon näiden luokkien vaikutukset mahdollisesti kumoavat toisensa ja näin ollen estimaatin suuruus saattaa pysyä saman tasoisena kuin alkuperäisessä aineistossakin. Vanhemmissa ikäluokissa vastauskadon määrä on vähäisempää. Ennustamista helpottaisi tieto koulutusryhmien jakautumi-

sesta ikäryhmittäin.

Koko aineiston keskiarvo pysynee lähellä alkuperäistä. Ikäluokittain oletan sukupuolten välisten erojen vaikuttavan tuloksiin, koska vastaajaluovuttajia ei ole rajattu sukupuolittain. Naisten matalammat arvot aineistossa tuovat miesten estimaatteja pienemmiksi ja vastaavasti miesten aineiston korkeat arvot tuottavat naisille korkeampia estimaatteja ikäryhmissä. Keskivirheiden oletan pienenevän hiukan.

Mallin 4 hypoteesi: mallin ikäryhmien estimaatit pysyvät alkuperäisen aineiston kaltaisina. Keskivirheet pienenevät hieman.

Malli 5: Tässä mallissa vastaajaluovuttaja-imputointi perustuu taustatietoihin sukupuolesta, ikäluokasta ja koulutuksesta. Taulukoiden 3.1 ja 3.3 perusteella nämä muuttujat lisäävät jokainen oman ulottuvuutensa luottamusindikaattorin arvoon. Arvo on suurempi miehillä kuin naisilla, kasvaa koulutuksen mukana ja toisaalta laskee ikäluokkien vanhetessa. Näistä muodostuu varmastikin hyvin kattavat kombinaatiot vastaajaluovuttajaryhmiin, joten oletan, että tämä malli tuottaa varsinkin sukupuolittain ikäryhmissä hyvin alkuperäistä aineistoa vastaavat estimaatit. Keskivirheiden oletan pienenevän havaintojen määrän kasvun myötä.

Mallin 5 hypoteesi: mallin ikäryhmien estimaatit pysyvät alkuperäisen aineiston kaltaisina, mutta koko aineiston estimaatti tulee kasvamaan hieman. Keskivirheet pienenevät hieman.

4.4 Imputointimallien estimaatit ja keskivirheet

Taulukossa 4.2 esitetään viiden imputointimallin tulokset alkuperäisen aineiston estimaattien kanssa. Taulukosta huomataan, että moni-imputointi tuottaa pienempiä keskivirheitä alkuperäiseen dataan verrattuna poikkeuksetta.

Taulukko 4.2
Luottamusindikaattorin estimaatit imputointimalleittain

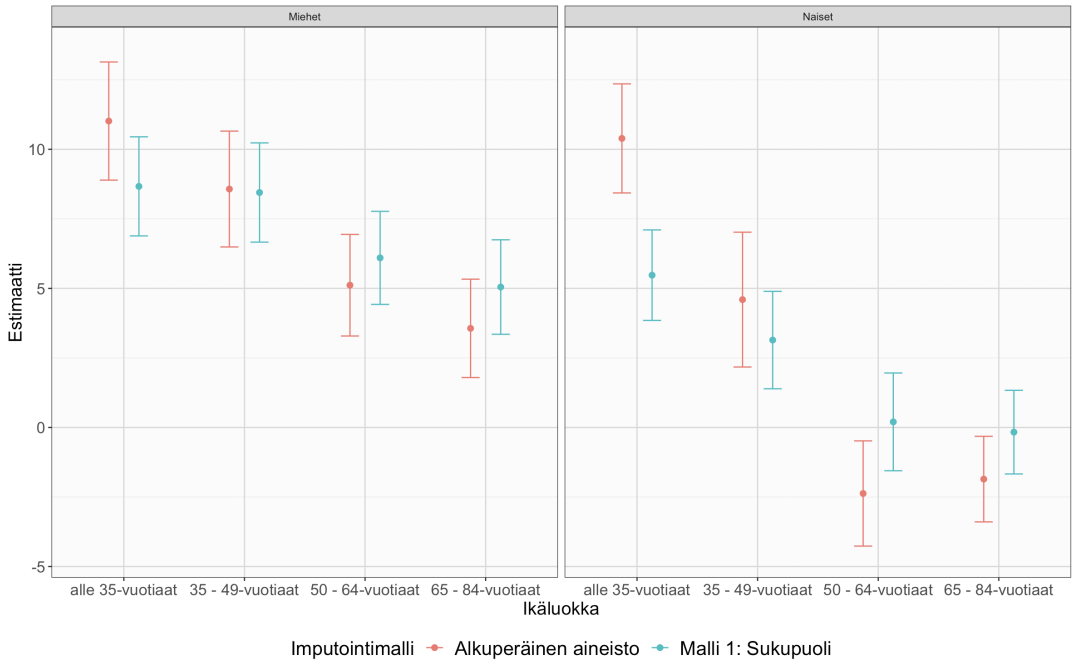
		Aineisto	Malli 1	Malli 2	Malli 3	Malli 4	Malli 5
		Estimaatti					
Koko aineisto		5,137	4,724	5,373	4,411	5,251	5,306
Miehet		7,494	7,239	7,527	6,767	6,433	7,575
Ikäluokka	alle 35	11,016	8,668	10,751	8,102	11,095	11,338
	35-49	8,572	8,446	8,755	7,794	7,876	8,185
	50-64	5,114	6,097	5,091	5,719	2,722	5,416
	yli 65	3,563	5,048	4,112	4,812	2,111	3,695
Naiset		2,824	2,255	3,260	2,099	4,090	3,080
Ikäluokka	alle 35	10,392	5,475	10,515	5,267	10,909	10,448
	35-49	4,597	3,142	4,497	3,311	5,676	4,884
	50-64	-2,373	0,201	-1,349	-0,227	-0,200	-2,209
	yli 65	-1,857	-0,169	-1,547	-0,284	-0,847	-1,680
		Keskivirhe					
Koko aineisto		0,716	0,589	0,619	0,602	0,639	0,593
Miehet		1,020	0,923	0,916	0,825	0,884	0,807
Ikäluokka	alle 35	2,124	1,781	1,730	1,617	1,771	1,691
	35-49	2,082	1,784	1,785	1,827	2,009	1,672
	50-64	1,826	1,673	1,676	1,696	1,630	1,806
	yli 65	1,768	1,698	1,611	1,753	1,664	1,586
Naiset		0,999	0,777	0,806	0,843	0,905	0,821
Ikäluokka	alle 35	1,961	1,627	1,500	1,556	1,623	1,622
	35-49	2,423	1,751	1,881	1,928	1,933	1,958
	50-64	1,891	1,757	1,704	1,692	1,633	1,584
	yli 65	1,538	1,503	1,323	1,563	1,484	1,359
		Variaatiokerroin					
Koko aineisto		0,139	0,125	0,115	0,137	0,122	0,112
Miehet		0,136	0,127	0,122	0,122	0,137	0,107
Ikäluokka	alle 35	0,193	0,205	0,161	0,200	0,160	0,149
	35-49	0,243	0,211	0,204	0,234	0,255	0,204
	50-64	0,357	0,274	0,329	0,297	0,599	0,334
	yli 65	0,496	0,336	0,392	0,364	0,788	0,429
Naiset		0,354	0,345	0,247	0,402	0,221	0,267
Ikäluokka	alle 35	0,189	0,297	0,143	0,295	0,149	0,155
	35-49	0,527	0,557	0,418	0,582	0,341	0,401
	50-64	-0,797	8,742	-1,264	-7,458	-8,147	-0,717
	yli 65	-0,828	-8,902	-0,855	-5,505	-1,752	-0,809

Malli 1: Sukupuoli
Malli 2: Sukupuoli ja ikäluokka
Malli 3: Sukupuoli ja koulutus
Malli 4: Ikäluokka ja koulutus
Malli 5: Sukupuoli, ikäluokka ja koulutus

Vertaillaan mallien hypoteeseja toteutuneisiin estimaatteihin:

Mallin 1 hypoteesi: malli siirtää jokaisen ikäryhmän estimaattia kohti sukupuolen keskiarvoa ja koko aineiston estimaatti tulee pienenemään. Keskivirheet pienenevät.

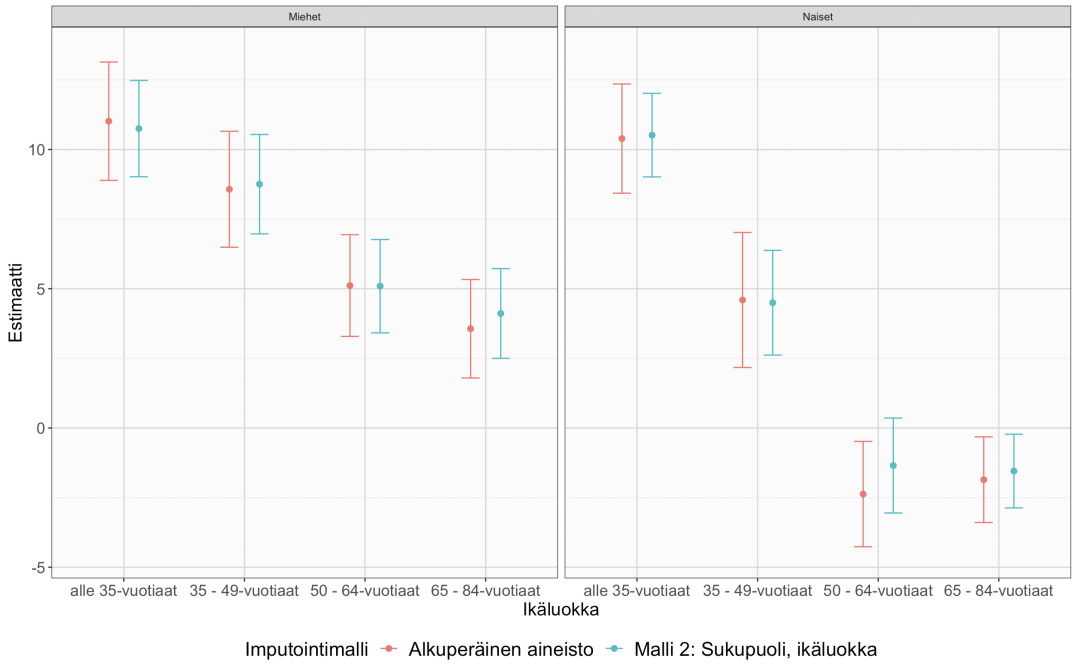
Mallin 1 koko aineiston estimaatti on 4,7, mikä on pienempi kuin alkuperäisen aineiston estimaatti 5,1. Miesten keskiarvo alkuperäisessä aineistossa on 7,5 ja naisten 2,8. Molemmissa sukupuolissa kaikkien ikäluokkien estimaatit estimoituivat alkuperäisestä arvostaan kohti sukupuolensa keskiarvoa (kuva 4.1).



Kuva 4.1: Imputointimalli 1: Luottamusindikaattorin estimaatit ja keskivirheet

Mallin 2 hypoteesi: mallin ikäryhmien estimaatit pysyvät alkuperäisen aineiston kaltaisina sukupuolittain, mutta koko aineiston estimaatti tulee kasvamaan hieman. Keskivirheet pienenevät hieman.

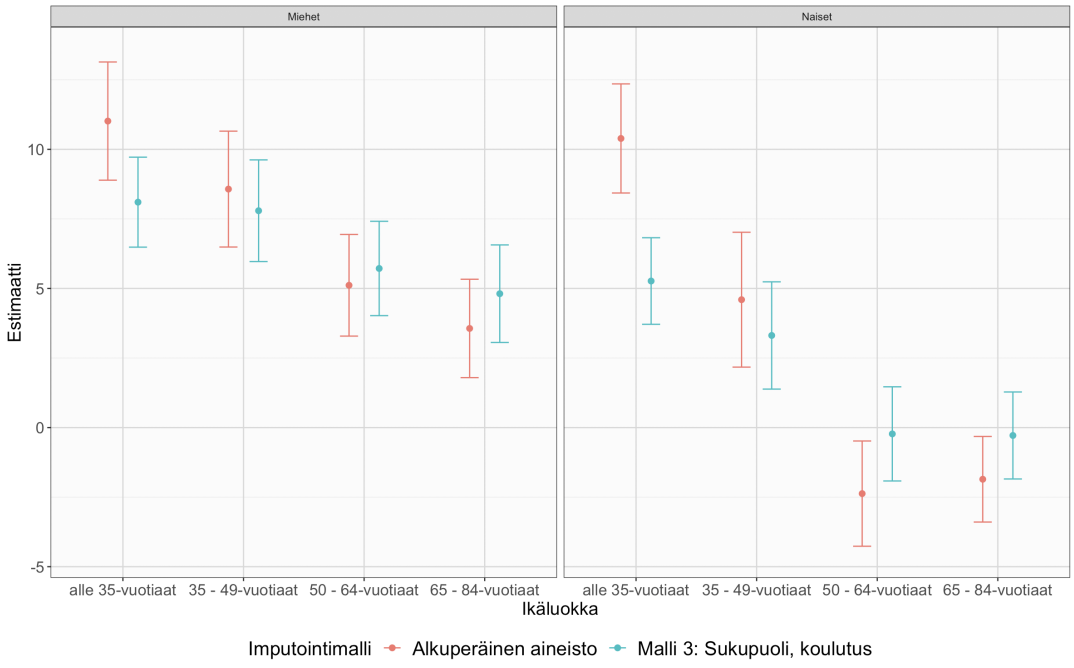
Hypoteesi näyttää toteutuneen aika tarkasti. Koko aineiston estimaatti estimoituu korkeammalle (5,3) kuin alkuperäisen aineiston estimaatti (5,1). Sukupuolittain ikäryhmätasolla estimaatit ovat hyvin lähellä alkuperäisen aineiston estimaatteja. Vanhimmissa ikäluokissa esiintyy eniten eroja alkuperäiseen estimaattiin verrattuna (kuva 4.2).



Kuva 4.2: Imputointimalli 2: Luottamusindikaattorin estimaatit ja keskivirheet

Mallin 3 hypoteesi: mallin koulutusluokkien estimaatit pysyvät alkuperäisen aineiston kaltaisina sukupuolittain, mutta koko aineiston estimaatti tulee pienenemään hieman. Keskivirheet pienenevät hieman.

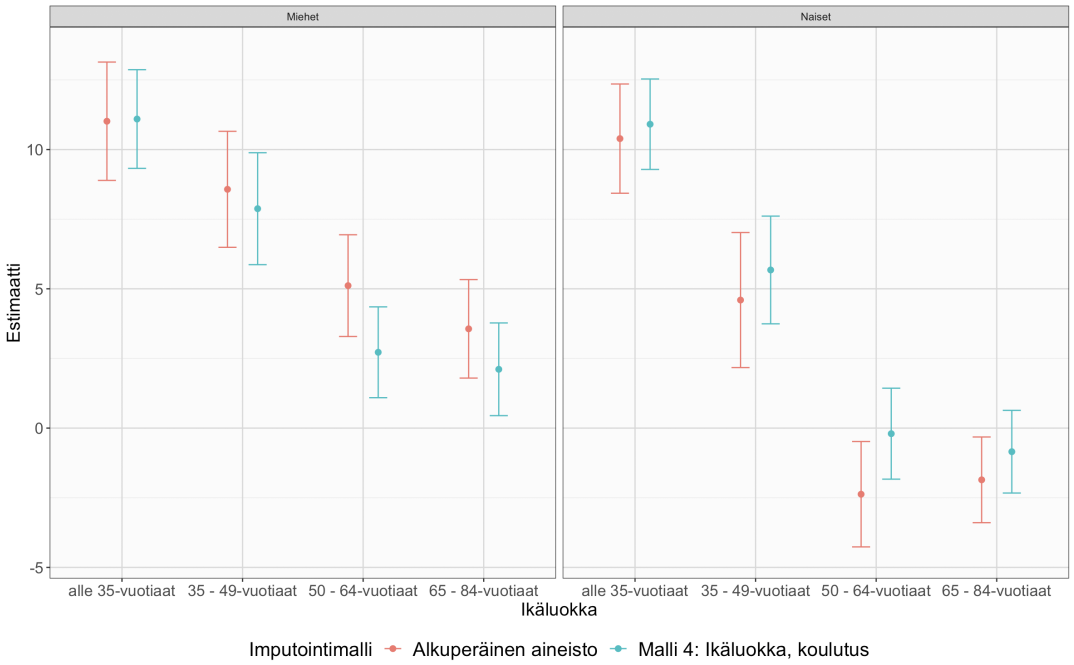
Tämän mallin hypoteesi ei ollut kovin tarkka. Koko aineiston luottamusindikaattorin estimaatti pieneni alkuperäisen aineiston estimaattiin verrattuna. Tämän mallin estimaatti on kaikista malleista pienin. Kuten mallissa 1, myös tässä mallissa vaikuttaa ilmenevän ikäluokkien arvojen estimoitumista kohti sukupuolen keskiarvoa (kuva 4.3).



Kuva 4.3: Imputointimalli 3: Luottamusindikaattorin estimaatit ja keskivirheet

Mallin 4 hypoteesi: mallin ikäryhmien estimaatit pysyvät alkuperäisen aineiston kaltaisina. Keskivirheet pienenevät hieman.

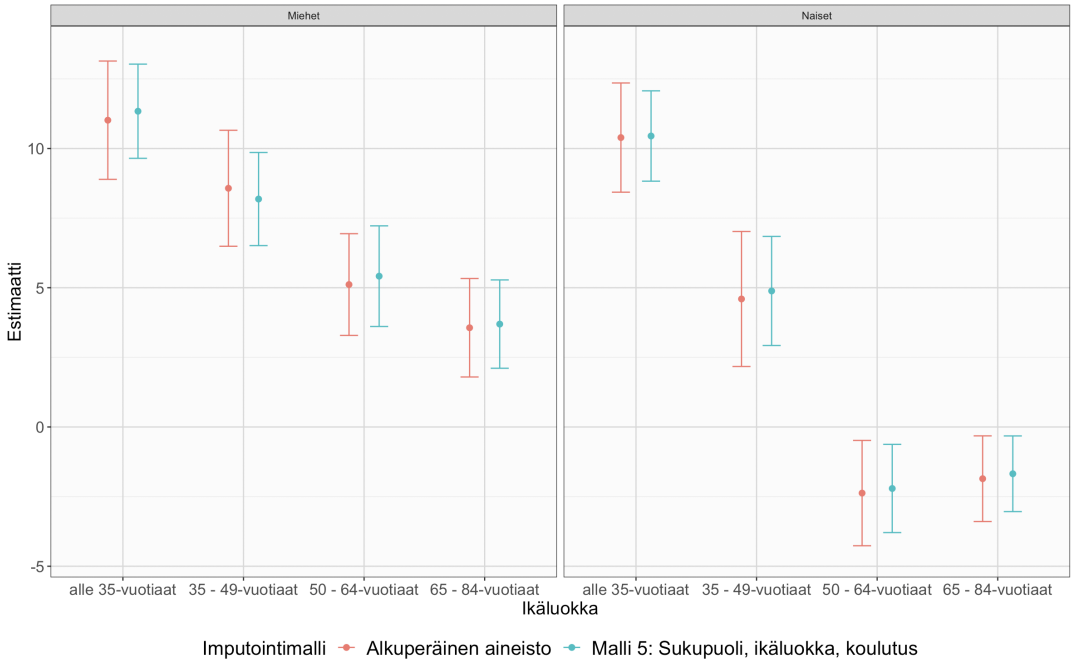
Tämän mallin ennustaminen oli hankalaa ja sen hypoteesi ei onnistunut. Koko aineiston estimaatti tosin osuu lähelle alkuperäien aineiston estimaattia. Sukupuolimuuttujan poissaolo näkyy tuloksissa siinä, että sukupuolen estimaatti, miehillä 6,7 (alkupäisessä aineistossa 7,5) ja naisilla 4,1 (alkupäisessä aineistossa 2,8) hakeutuu kohti aineiston keskiarvoa 5,1 (kuva 4.4).



Kuva 4.4: Imputointimalli 4: Luottamusindikaattorin estimaatit ja keskivirheet

Mallin 5 hypoteesi: mallin ikäryhmien estimaatit pysyvät alkuperäisen aineiston kaltaisina, mutta koko aineiston estimaatti tulee kasvamaan hieman. Keskivirheet pienenevät hieman.

Tämä hypoteesi onnistui aika hyvin ennustamaan mitä mallissa tapahtui. Taulukosta 4.2 voidaan nähdä, että kaikki estimaatit osuvat hyvin lähelle alkuperäisen aineiston arvoja (kuva 4.5).

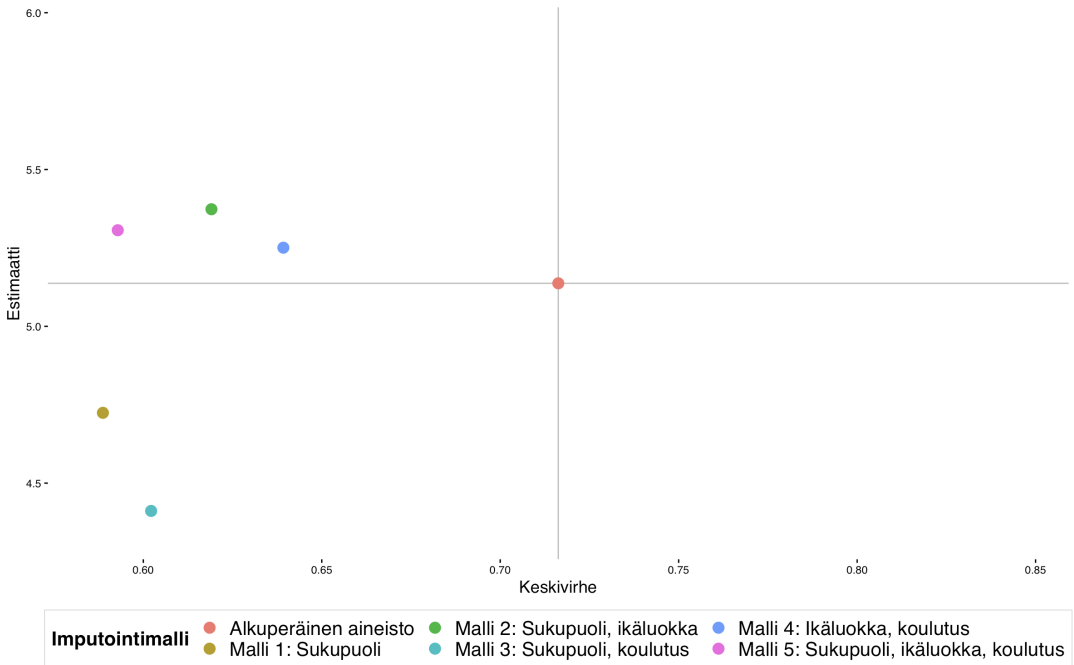


Kuva 4.5: Imputointimalli 5: Luottamusindikaattorin estimaatit ja keskivirheet

5 Tulokset ja pohdinta

Edellisessä luvussa asetettiin imputointimallien estimaateille hypoteeseja taustamuuttujien perusteella. Tuloksista huomattiin, että mallissa mukana olevien taustamuuttujien perusteella oli mahdollista asettaa hypoteeseja oikeaan suuntaan tarkastelemalla luottamusindikaattorin estimaatin arvoja sukupuolien, ikäluokkien ja koulutusluokkien välillä alkuperäisessä aineistossa sekä samoissa luokissa esiintyvän vastauskadon määrää.

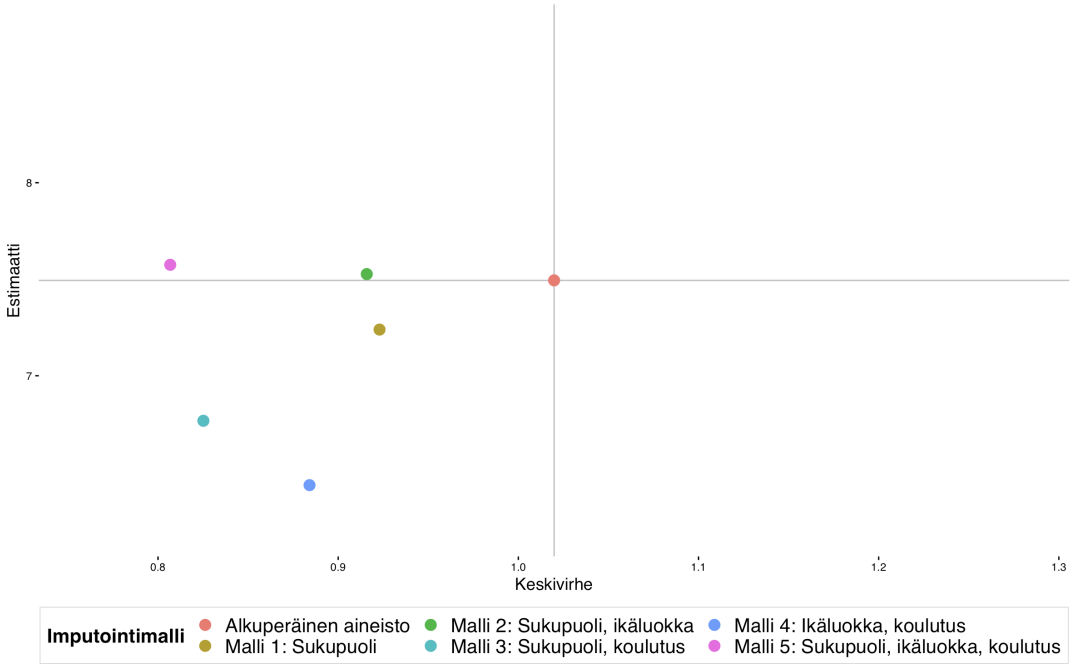
Sijoittamalla imputointimallien tulokset nelikenttään, voidaan visuaalisesti tulkita vastauskadon vaikutuksia luottamusindikaattorin arvoon. Nelikentän keskipisteessä on alkuperäisen vastauskatoa sisältävän aineiston estimaatti. Y-akseli määrittää estimaatin arvon ja x-akseli estimaatin keskivirheen. Tulostamalla samaan kuvaajaan imputointimallien tuottamat estimaatit, nähdään oliko mallin estimaatti isompi vai pienempi kuin alkuperäisen aineiston estimaatti ja oliko mallin keskivirhe suurempi vai pienempi kuin alkuperäisen datan keskivirhe (kuva 5.1).



Kuva 5.1: Imputointimallien luottamusindikaattorin estimaatit ja keskivirheet

Kuvasta 5.1 nähdään, että kaikkien mallien tuottamien moni-imputoitujen aineistojen keskivirheet ovat pienempiä kuin alkuperäisen aineiston keskivirhe. Tätä voi selittää sillä, että puuttuvien arvojen korvaaminen tapahtuu aidoilla alkuperäisessä datassa jo olevilla arvoilla. Seurauksena havaintoyksiköiden määrä kasvaa ja hajonta pienenee.

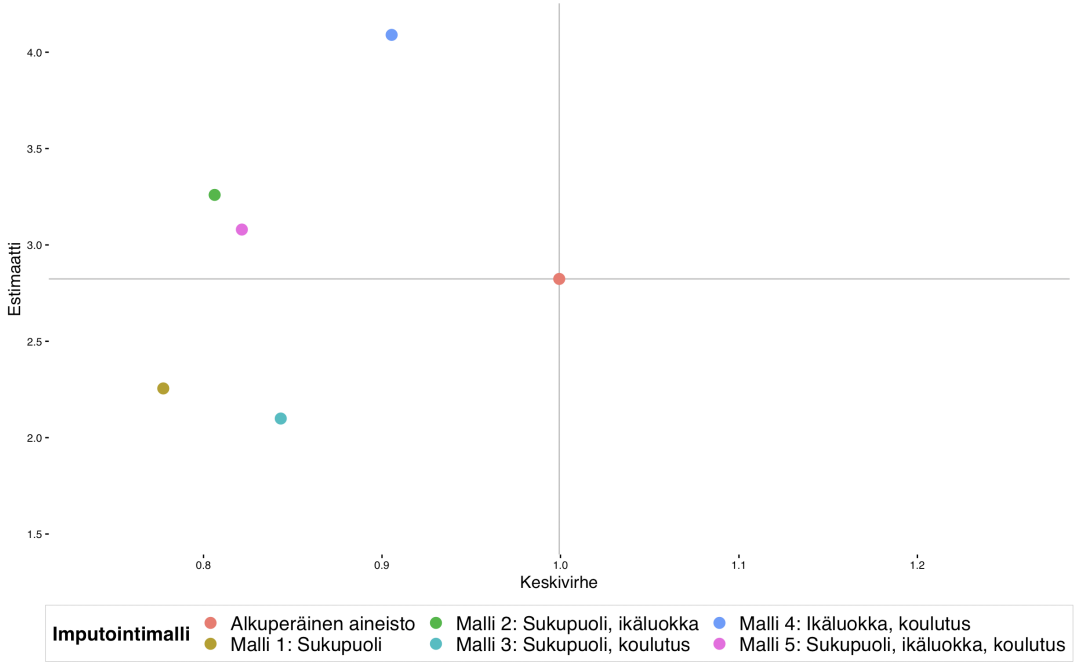
Imputointimallit 1 ja 3 ovat malleja, joissa ei ole ikäluokkatietoa. Niiden tuottamat luottamusindikaattorin arvot estimoituvat pienemmäksi kuin alkuperäisen aineiston arvo. Imputointimalleissa 2, 4 ja 5, joissa ikäluokkatieto on mukana,luottamusindikaattorin arvo estimoituu suuremmaksi kuin alkuperäisessä aineistossa. Ikäluokka vaikuttaisi siis määrittävän estimaatin suuruutta, kun tarkastellaan aineistoa ilman sukupuoli- ja ikäluokkajaottelua.



Kuva 5.2: Imputointimallien luottamusindikaattorin estimaatit ja keskivirheet miesten aineistoissa

Tarkastellaan seuraavaksi mallien estimaatteja myös sukupuolittain. Kuvassa 5.2 miesten aineistojen estimaateista huomataan, että mallien 2 ja 5 estimaatit ovat hyvin lähellä alkuperäisen aineiston estimaatin arvoa. Mallin 5 lisätieto koulutuksesta pienentää sen keskivirhettä verrattuna malliin 2, joka perustuu

tietoon sukupuolesta ja ikäluokasta. Mallin 4 estimaatti on kaikista malleista pienin, mikä johtuu siitä, että mallissa ei ole mukana sukupuolitietoa. Mallin 3 taustatiedot sukupuoli ja koulutus tuottavat miesten aineistossa kauas aineiston estimaatista sijoittuvan arvon. Tieto ikäluokasta vaikuttaa olennaiselta tässäkin tapauksessa. Mallin 1, joka perustuu ainoastaan sukupuolitietoon, estimaatti sijoittuu hyvin lähelle alkuperäisen aineiston arvoa.



Kuva 5.3: Imputointimallien luottamusindikaattorin estimaatit ja keskivirheet naisten aineistoissa

Naisten aineistossa mallin 1 estimaatti sen sijaan ei sijoitu lähelle alkuperäisen aineiston arvoa, vaan arvoltaan pienemmäksi (kuva 5.3). Tässäkin aineistossa mallien 2 ja 5 estimaatit ovat hyvin lähellä alkuperäisen aineiston estimaatin arvoa, mutta naisten aineistossa mallin 5 lisätieto koulutuksesta ei pienennä estimaatin keskivirhettä verrattuna malliin 2. Mallin 4 estimaatti on naisten aineistossa suurin, mikä johtuu tässäkin aineistossa siitä, että mallissa ei ole mukana sukupuolitietoa, jolloin estimaatti sijoittuu lähemmäksi koko aineiston estimaattia.

Mallin 3 taustatiedot sukupuoli ja koulutus tuottavat naistenkin aineistos-

sa kauas aineiston estimaatista sijoittuvan arvon. Tieto ikäluokasta vaikuttaa olennaiselta tässäkin aineistossa. Kun malli ei ota huomioon ikäluokkaa, niin vastaajaluovuttajien enemmistö koostuu vanhimmista ikäluokista, joiden luottamusindikaattorin arvot puolestaan ovat molemmissa sukupuolissa alhaisimmat. Puuttuvia tietoja paikataan suhteessa enemmän alhaisilla arvoilla, ja mallin estimaatti estimoituu pienemmäksi.

Taustamuuttujatietoja ei aina ole saatavilla yhtä kattavasti kuin Tilastokeskuksen väestötietokannasta kerätyissä otoksissa. Ilman kattavia taustatietoja puuttuneiksi jääneistä yksiköistä imputointimallin tulokset saattavat vääristää tulkittavia tuloksia. Mallien tuloksia tulkittaessa voidaan malleja lähestyä ajatuksella tilanteesta, jossa kyseisen mallin taustamuuttujat olisivatkin ainoat saatavilla olleet taustamuuttujat. Lisäksi tuloksia tulkittaessa on hyvä tiedostaa, että kyselytutkimuksen aihe saattaa olla yhteydessä todennäköisyyteen osallistua kyselyyn (Groves, Presser & Dipko, 2004). Tässä tapauksessa voidaan ajatella, että koska kyseessä on kuluttajien luottamusbarometri, niin hyvätuloisilla ja talousasioista kiinnostuneilla henkilöillä voi esiintyä suurempaa kiinnostusta myös tähän kyselyyn vastaamisessa. Toisaalta, kiinnostus talousasioihin ei välttämättä ole yhteydessä luottamusindikaattorin arvoon.

Koko aineiston estimaatteja vertaillessa mallien tuottamien estimaattien välillä on eroja, mutta erot eivät ole käytännössä kovin suuria. Imputointimallien estimaatit koko aineiston tasolla vaihtelevat välillä 4,411 - 5,373. Luottamusindikaattori voi saada arvoja välillä (-100, +100) ja kuukausien välillä arvo vaihtelee tavallisesti jopa useita kokonaislukuja (Tilastokeskus, 2020).

6 Johtopäätökset

Tämän tutkielman tärkein tavoite oli selvittää vastauskadon vaikutusta kuluttajien luottamusindikaattorin arvoon. Aineisto kärsii muiden nykyaikaisten survey-aineistojen tapaan isosta vastauskadon osuudesta erityisesti nuorimmissa ikäluokissa. Vastauskatoa paikattiin moni-imputoimalla käyttäen vastaajaluovuttaja-menetelmää.

Vertailtavaksi luotiin viisi imputointimallia käyttäen vaihtelevia yhdistelmiä taustamuuttujista. Malleihin hyväksytyt taustamuuttujat olivat tilastollisesti merkitsevässä yhteydessä kuluttajien luottamusindikaattorin muodostavien muuttujien vastauskatoon ja kuluttajien luottamusindikaattorin arvoon. Poikkeuksena tästä oli vastaajan sukupuoli, joka ei ollut yhteydessä vastauskatoon. Tuloksia tarkasteltiin aineiston kokonaistasolla sekä ikäluokissa sukupuolittain.

Jokainen malli pienensi estimaattien keskivirheitä kaikissa tarkasteluryhmissä verrattuna alkuperäisen aineiston keskivirheisiin. Luottamusindikaattorin arvot eroavat sukupuolten ja ikäluokkien välillä alkuperäisessä aineistossa. Selkeimmän eron mallien välillä tuotti ikäluokkatiedon puuttuminen mallista. Ikäluokkamuuttujan puuttuminen mallista pienensi luottamusindikaattorin estimaattia alkuperäiseen arvoon verrattuna sekä koko aineiston tasolla että sukupuoliryhmissä.

Parhaiten alkuperäistä aineistoa vastaavat luottamusindikaattorin estimaatit tuotti malli, jossa vastaajaluovuttajaluokkia oli eniten (sukupuoli, ikäluokka, koulutus). Luokkatietojen lisääminen malliin tarkentaa vastaajaluovuttajien ryhmää ja samalla myös asettaa haasteen vastaajaluovuttajien lukumäärän suhteen. Tiettyjen luovuttajaryhmien vähäisen luovuttajamäärän vuoksi tässä tutkielmassa siviilisääty-muuttuja rajattiin tarkastelun ulkopuolelle. Jos aineisto olisi isompi, niin siviilisäätymuuttuja olisi voinut olla hyödyllinen lisämuuttuja mallien vertailuissa.

Tämän tutkielman tarkastelujen ulkopuolelle rajattiin tieto vastaajan tuloista. Tulevissa tarkasteluissa olisi mielenkiintoista yhdistää myös tämä taustatieto mukaan imputointimalliin. Tuloluokan tapauksessa ongelmaksi saattaa myös muodostua malliluovuttajaluokkien pienentyminen liian pieneksi.

Lähteet

Alwin, D. F. (2007). Margins of error: A study of reliability in survey measurement. Hoboken, N.J.: Wiley-Interscience.

Durrant, G. (2005). Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review. National Centre for Research Methods Working Paper Series, June 2005.

Little, R. J. A. & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). Hoboken, N.J.: John Wiley.

Laaksonen, Seppo (2013). Surveymetodiikka. Aineiston kokoamisesta puhdistamisen kautta analyysiin. bookboon.com. Saantitapa: `https://bookboon.com/fi/surveymetodiikka-ebook`

Rubin, D. B. (2004). Multiple imputation for nonresponse in surveys. Hoboken, N.J.: Wiley-Interscience.

van Buuren, S. (2018). Flexible Imputation of Missing Data, Second Edition. Chapman and Hall/CRC.

Suomen virallinen tilasto (SVT): Kuluttajien luottamus [verkkojulkaisu]. ISSN=2669-8862. Helsinki: Tilastokeskus [viitattu: 26.4.2020].
Saantitapa: `http://www.stat.fi/til/kbar/2017/08/kbar_2017_08_2017-08-28_laa_001_fi.html`

van Buuren, S., Boshuizen, H.C., and Knook, D.L.(1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. "Statistics in Medicine,18

Allison, P. D., (2005). 113-30: Imputation of Categorical Variables with PROC MI.
Saantitapa: `https://support.sas.com/resources/papers/proceedings/proceedings/sugi30/113-30.pdf`

Groves, R. M., Presser, S., Dipko, S. (2004). The Role of Topic Interest in Survey Participation Decisions. Public Opinion Quarterly, Volume 68, Issue 1, March 2004, Pages 2–31.

Suomen virallinen tilasto (SVT): Kuluttajien luottamus [Findikaattori.fi]. ISSN=2669-8862. Helsinki: Tilastokeskus [viitattu: 18.5.2020].
Saantitapa: `https://findikaattori.fi/fi/table/104`

Liitteet

Liite 1: Moni-imputoinnin vaiheet SAS -koodilla

```
*       1. Imputointivaihe: HOT-DECK-imputointi,
           mallina sukupuoli/ikaluokka/koulutus;

proc surveyimpute data=kbm_tammi method=hotdeck(selection=weighted)
                  seed=54321 ndonors=20;
    var k2 k8 k9 k10b;
    cells sp ikayks tutk;
    id kohdenro;
    weight wcal2;
    output out = sp_ika_tutk donorid;
run;

* koostetaan ja numeroidaan imputoidut datasetit ;

data imp;
set sp_ika_tutk;
if (ImpIndex = 0) then do;      /* sisällytetään kokonaiset rivit        */
    do _Imputation_ = 1 to 20;  /* mukaan kaikkiin imputointiaineistoihin */
        output;
    end;
    end;
else do;
    _Imputation_ = ImpIndex;
        output;
    end;

proc sort data=imp;
    by _Imputation_ kohdeno;
run;
```

Liite 2: Moni-imputoinnin vaiheet SAS -koodilla

```
* Luottamusindikaattorin laskeminen moni-imputoituun dataan ;

data imp_cci;
set imp;
*Q1: Oma taloudellinen tilanne nyt;
q1=100*(1*(k8=1)+.5*(k8=2)-.5*(k8=4)-1*(k8=5));
*Q2: Oma taloudellinen tilanne 12 kk päästä;
q2=100*(1*(k9=1)+.5*(k9=2)-.5*(k9=4)-1*(k9=5));
*Q4: Suomen talous 12 kk päästä;
q4=100*(1*(k2=1)+.5*(k2=2)-.5*(k2=4)-1*(k2=5));
*Q9: Rahankäyttö kestokulutukseen 12 kk päästä;
q9=100*(1*(k10b=1)+.5*(k10b=2)-.5*(k10b=4)-1*(k10b=5));
new_cci=(q1+q2+q4+q9)/4;
run;

* 2. Analyysivaihe: lasketaan luottamusindikaattorin estimaatit aineistoihin ;

    ods output Domain=outsummary;
proc surveymeans data=imp_cci;
    domain _imputation_;
    var new_cci;
    weight wcal2;
run;
    ods output close;

* 3. Yhdistämisvaihe: yhdistetään imputoitujen aineistojen tulokset ;
proc mianalyze data=outsummary;
modeleffects mean;
stderr StdErr;
run;
```